Research Statement

Kenneth Mai
Stanford University

Summary

My primary research interests lie at the overlap of computer architecture and VLSI circuit design. In the future, with the underlying implementation technology increasingly affecting architecture and circuit design, we must adapt and reinvent current designs to circumvent technology constraints and to target emerging applications. The key near-term challenge is to build compute systems that can efficiently achieve high-performance, yet remain economically feasible, general-purpose, and easy to program. I am interested in designing and implementing both reconfigurable and hardwired solutions to address this need. In the long-term, with CMOS scaling approaching fundamental limits, the challenge is to build compute systems from technology building blocks that may be radically different from those we use today. With my experience in both architecture and circuits, I feel well-placed to recognize how to exploit new technologies for ever more efficient and high-performance compute systems.

Current work

Conventional microprocessor designs are running out of steam. Trends in VLSI process technology, such as growing interconnect delay, restrictive power constraints, and increasing device leakage, make monolithic processor designs increasingly difficult and inefficient to scale up. Yet these advanced processes also offer us an unprecedented amount of die area to fill. Even if we could easily extend the architectures, traditional single-threaded applications have limited instruction-level parallelism, most of which is already mined out by current machines. However, decades of Moore's Law and the toil of countless engineers have opened up new application areas, such as computational biology and ubiquitous sensor networks, where previously compute was either not powerful enough or not cheap enough to attack the problem. These new applications will keep up the demand for ever more efficient, high-performance computation previously only available from custom ASICs. However, increasing design cost and complexity require that new designs be high-volume and general-purpose to remain economically feasible. So the key challenge for the near-term future is to build a processor that has the efficiency and performance close to that of custom ASICs, yet still retains the generality and programmability of today's CPUs.

One approach to this challenge is to use reconfigurable logic. With FPGAs, researchers have achieved orders-of-magnitude better performance and efficiency over general-purpose processors on some applications. However, for many applications the fine-grained reconfigurability of FPGAs is unnecessary, and they suffer from large overheads. The Stanford Smart Memories project [1] explores coarse-grain reconfigurable processors seeking to remain flexible while incurring low overheads. A Smart Memories chip is made up of multiple modular reconfigurable processing tiles, each containing local memory, interconnect, a processor core, and a network interface. We have taken the design from the architectural specification and simulation of a group of tiles [1] down to the circuit-level implementation of the reconfigurable memory blocks [2] and high-speed, low-power interconnect [3].

I was principally responsible for the design and implementation of the reconfigurable memory system. Previous reconfigurable memory systems have offered flexibility at the extremely low level (*e.g.* memory from FPGA LUTs) or at the extremely high level (*e.g.* cache way/bank shutdown). By leveraging the partitioned nature of large memories [4] and adding the minimum set of necessary features to each partition, we designed a low-overhead, flexible memory system that could emulate various cache configurations, streaming FIFOs, and scratch-pad memories. After specifying the memory system architecture down to the HDL level, I led a team of five graduate students to design and fabricate a testchip containing prototype reconfigurable memories on a 9.9mm^2 die in a TSMC 0.18μm 6-metal CMOS process. We functionally verified the chip at 1.1GHz, 1.8V Vdd, and room temperature this past summer [2].

Future Work

My research experience at Stanford has piqued my interest in three main areas:

- Next-generation general-purpose processor architectures
- Memory system design
- CAD tools for visualization and verification

For next-generation processor designs, we have an unprecedented amount of die area to fill, but power constraints limit the amount of logic we can run at any one time. Under these conditions, heterogeneous multi-core architectures present an intriguing alternative to reconfigurable logic for efficient, high-performance compute. These designs have multiple different processor cores on die, each hardwired for a specific class of applications. The cores that best suit the application are run at full Vdd and clock speed, while other less useful cores run at a reduced Vdd and clock or are completely shutdown to minimize static power dissipation. By steering power to the cores best suited to run the application, we achieve maximum power efficiency and performance. The die as a whole remains general purpose, since there are multiple cores that can efficiently execute a wide range of applications.

As a first step to designing such a system, I would undertake a comprehensive study of the energy efficiency of modern processor and ASIC designs for a broad set of applications to determine the optimal computation paradigm for each application. Using the study and VLSI constraints to guide the types of cores to include, I could then embark on the design of a prototype system. In addition to the architectural questions, interesting circuit challenges may arise in the design of the core interfaces, power grid, and clock distribution network.

Whatever the shape of future compute, on-die memory will continue to occupy a large percentage of the processor die area, account for much of the power, and have a large impact on performance. I would like to continue and extend my work on reconfigurable memories beyond general-purpose computing to high-volume ASIC application areas such as graphics and networking. Additionally, there are multiple opportunities in pursuing custom high-speed, high access rate memory designs for networking search structures (*e.g.* ternary CAMs) and in high-speed I/O.

While much of memory design has been custom VLSI work, the field is moving towards automated generation. Leveraging work done at Stanford [4], I want to build a freeware compiler for self-timed, low-power memories and caches. This tool would not only meet the needs of chip designers for fast on-die memory, but also serve as an extremely accurate memory/cache modeler. Existing memory/cache modelers use a number of simplifying assumptions that make them unsuitable for bleeding-edge memory evaluation and can generate inaccurate unrealistic results. This work could serve as the starting point for an open source library of ready-made processor building blocks and generation tools, similar to the HDL models from the Opencores initiative [5], but at a schematic and layout level. This library could speed testchip design and promote standardization and design reuse in the academic architecture and VLSI design communities.

In addition to memory design tools, I am interested in developing CAD tools for circuit simulation visualization and automated verification. An enhanced schematic capture or layout tool that overlays and "plays" simulation results over the design, similar to the IBM PICA movies [6], would help the designer understand the circuit operation and speed verification. While this type of visualization tool is useful, relying solely on the overworked, sleep-deprived designer/graduate student to catch mistakes is not foolproof. Automated simulation-based circuit checkers, similar to STAR [7] but for digital circuits, would aid in catching bugs and increase design portability.

For the long-term, the enduring challenge is that underlying implementation technologies will continue to affect processor architecture and circuit design. As process technologies edge ever closer to fundamental scaling limits, device characteristics will change even more, perhaps prompting a jump to new device technologies such as fin-fets or carbon nanotubes. Designers must continually adapt their architectures and circuits to the vagaries of the implementation technologies and to the demands of the applications to perform feasible and relevant work. This necessitates forming strong collaborative bonds and embarking on multi-disciplinary research projects. With the coming changes in device technologies and the emergence of exciting new applications, architecture and circuit design will remain rich and vibrant fields of research.

References

[1] K. Mai, *et al*., "Smart memories: a modular, reconfigurable architecture," *International Symposium on Computer Architecture*, June 2000.

[2] K. Mai, *et al.*, "Architecture and circuit techniques for a reconfigurable memory block," *International Solid-State Circuits Conference*, February 2004.

[3] R. Ho, *et al.* "Efficient on-chip global interconnects," *Symposium on VLSI Circuits*, June 2003.

[4] B. Amrutur, *et al*., "Speed and power scaling of SRAM's," *Journal Solid-State Circuits*, February 2000.

[5] http://www.opencores.org

[6] http://www.research.ibm.com/topics/popups/serious/chip/html/pica.html

[7] D. Liu, *et al.*, "A framework for designing reusable analog circuits," *International Conference on Computer Aided Design*, November 2003.